Title: Superstore Customer Behavior Analysis: Predicting Response to Gold Membership Offer – Final Report

Group 20 – MGT-6203:

- Benavides, Ramon
- Covarrubia, Rory
- Fabrega Tapia, Cesar
- Perez, Angie
- Sevilla, Alejandra

1. Choice of Topic, Business Justification and Problem Statement

1.1. Choice of Topic

Our selection of this topic was primarily driven by the recognition of the influence that strategic initiatives such as the launch of a Gold Membership can have on the operational functionality and success of superstores. The goal of these memberships is to create deeper customer engagement and loyalty, thereby increasing customer lifetime value. Considering the potential for an increase in revenue generation and customer retention that these memberships hold, it becomes critical to precisely target the customers who are more likely to accept this membership. In conclusion, we selected this topic, with the goal of using data-driven insights to improve the effectiveness of the Gold Membership initiative.

1.2. Business Justification

The introduction of the Gold Membership program in the superstore is strategically aligned with current market trends and consumer behavior. This observation is supported by the article prepared by McKinsey & Company in 2020, “Coping with the big switch: How paid loyalty programs can help bring consumers back to your brand” and an article by Katie Deighton, published in 2023, in the Wall Street Journal, “Want Better Customer Service? Join the (Membership) Club”, businesses that have adopted paid loyalty programs, including Meta, P.F. Chang’s, Greenlight, Booking.com, and Sony’s PlayStation, have experienced improved customer loyalty and secure cash flow. The insight from these success stories supports the potential efficacy of the Gold Membership program.

The Gold Membership program offers members a 20% discount on all product purchases for an annual fee of $499, significantly less than the standard $999. This strategy is designed to increase spending, offering value to customers while also providing a steady source of revenue for the superstore.

With the increasing inflation rates, the 20% discount offered by the Gold Membership program could be highly attractive to consumers. By accurately targeting the right customers, we expect to optimize resource usage, reduce campaign costs, and increase the initiative's success rate, thereby creating a win-win situation for both customers and the superstore.

In summary, the Gold Membership program initiative is not just a response to market changes but a strategic move to deepen customer engagement, increase customer loyalty, and enhance customer lifetime value for the Superstore.

1.3. Problem Statement

The main challenge that our project addresses involve the accurate identification of potential Gold Members. To solve this, our plan is to create a predictive model that can successfully identify these customers from the entire customer base. Such a model would provide the superstore with a tool to channel its promotional efforts more efficiently, thereby amplifying the return on its strategic initiative. An approach where this offer is promoted to the entire customer base can increase costs and potentially compromise return on investment.

Based on successful strategies employed by Meta and other well-known industry companies, as highlighted in the Wall Street Journal article, we recommend a more targeted strategy. These companies have proven the effectiveness of tailoring their membership programs to certain audience segments.

By adopting a similar approach, we aim to promote the Gold Membership initiative to target customers with the greatest likelihood of subscription. In doing so, we will maximize the campaign's potency and, by extension, the return on investment.

2. Methodology and General Approach

The methodology and approach taken is the following:

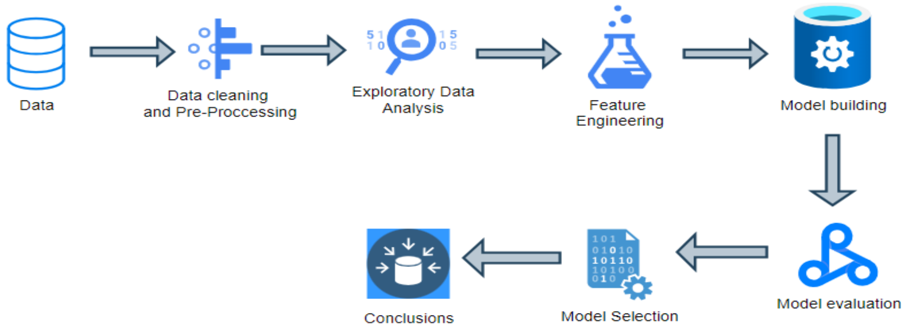

Figure 1 - Methodology and General Approach

1. **Data Cleaning and Pre-Processing**: this initial step involved preparing the data for analysis. We explored the structure of the data, handled missing values, and performed necessary data transformations. Categorical variables were identified and appropriately encoded, ensuring they could be effectively utilized in our predictive models.
2. **Exploratory Data Analysis:** we focused on gaining a better understanding of the data and identifying any underlying patterns or anomalies. We used histograms, bar plots, and other visualizations to examine the distributions and relationships of the variables. Outliers were addressed to prevent model bias, and a correlation matrix helped us understand the relationships between variables.
3. **Feature Engineering:** this phase involved creating new features to improve model performance. Some variables were converted to binary format, and interaction terms were created.
4. **Model Building:** we built and evaluated various predictive models, with a primary focus on logistic regression. Models were created with different sets of predictive variables, enabling us to compare and identify the most effective approach for predicting customer response.
5. **Model Evaluation:** the data was split into training and testing sets to avoid overfitting and test the model's performance on unseen data. We assessed the models using performance metrics like Area Under the ROC Curve (AUC-ROC), accuracy, specificity, and precision.
6. **Model Selection and Conclusions:** we compared the performance of all models developed, selected the most effective one, and drew conclusions. The chosen model provided us with insights into the characteristics of customers likely to purchase the Gold Membership.

3. Understanding of the Data and Exploratory Data Analysis

3.1. Dataset Source / Understanding the Data

1. **Dataset Source:** the dataset "superstore_data.csv" was obtained from Kaggle (Kaggle Link - Superstore Marketing Campaign Dataset), a platform for data science projects. To facilitate the collaboration between team members, the dataset was hosted on Dropbox (Dropbox Link - Dataset).
2. **Data Structure:** containing 2240 observations and 22 variables, the dataset offers a comprehensive representation of both customer demographic information and behavioral data. Variables go from demographic details such as the customer's age, education level, marital status, number of children, and household income, to behavioral data such as expenditure in different product categories, frequency of discounted purchases, number of website visits, and the recency of purchases.
3. **Preliminary Data Observations:** initial analysis provided multiple opportunities for data transformation. The 'Dt_Customer' variable can be converted from a character type to a date variable, while 'Education' and 'Marital_Status' can be converted to factors. Interaction variables can be introduced to capture complex relationships, such as the influence of 'Income' on 'NumStorePurchases' and the impact of 'Education' on 'Income'. Additionally, the individual spending amounts in different categories can be consolidated into a cumulative expenditure variable, offering a view of each customer's total spending.

3.2. Data Cleaning

During the data cleaning phase, we took various steps to ensure that our dataset was appropriate and ready for analysis. These steps were:

1. **Checking for Duplicates:** no duplicate values were found, so no deletion of rows based on duplication was required.
2. **Handling Missing Values:** we had two main options: substituting missing values with statistical estimates (e.g., mean, median, mode), or eliminating them if the number of missing values was relatively small. 24 ‘Income’ entries were missing. Given the wide range of the 'Income' variable, using the mean or median for imputation might not be representative, we chose to remove instead of substituting.
3. **Data Transformation:** we then applied some transforming steps to our data to facilitate our prediction model building. 'Dt_Customer' was transformed to a date type and 'Id' column was dropped for lack of significance. Expenditure subcategories were consolidated into a 'MntTotalSpent' column. 'Education' and 'Marital_Status' were converted to factors, non-standard entries in 'Marital_Status' removed, and 'Basic' and 'Single' set as reference levels. Dummy variables were created for categorical variables. After all these steps, the dataset was clean and ready for analysis and model building.

3.3. Exploratory Data Analysis

In our exploratory data analysis section, we performed the following actions:

1. **Histogram for Continuous Variables:** they provided insight into the distribution of continuous values in our dataset.
   - Income: the histogram showed most customers having an income between 0 and 100K, with a small number of outliers in the 150K range.
   - Year of Birth: concentration of birth years indicating that most customers are adults. However, we noticed a few outliers for customers born on or before 1900.
   - Recency, Number of Deals Purchases, Number of Web Purchases, Number of Catalog Purchases, Number of Store Purchases, Number of Web Visits per Month, Total Amount Spent: histograms for these variables were also constructed to analyze their distributions.

The outliers found could either be data entry errors or represent a small portion of high-income earners and older customers. We investigated them in the following sections.

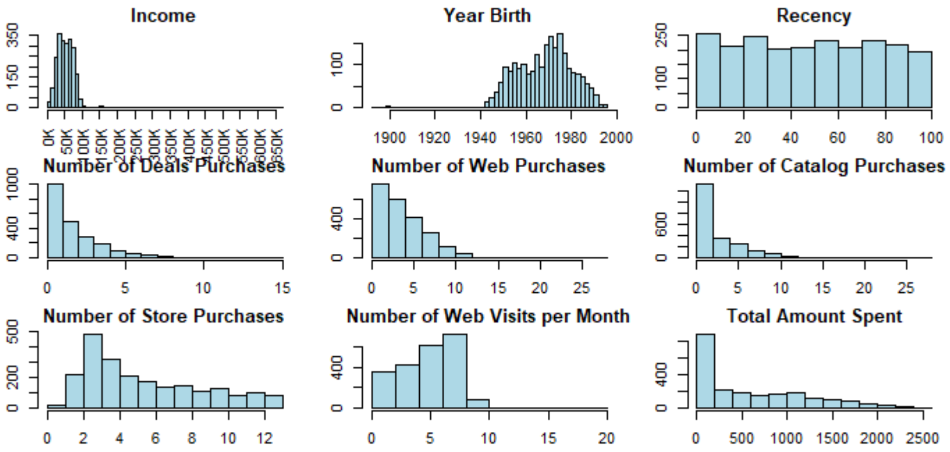

Figure 2 - Histograms of Continuous Variables

2. **Bar Plots for Categorical Variables:** the 'Education' bar plot indicated that most of our customers are graduates. The 'Marital_Status' bar plot showed that most customers are either 'Married' or 'Together'. For 'Kidhome' and 'Teenhome', we observed an approximate even split between customers who do and do not have kids or teenagers at home. However, the number of customers with two kids or teenagers is minimal.

   The 'Complain' and 'Response' bar plots showed that despite the relatively low number of complaints, most responses are negative.

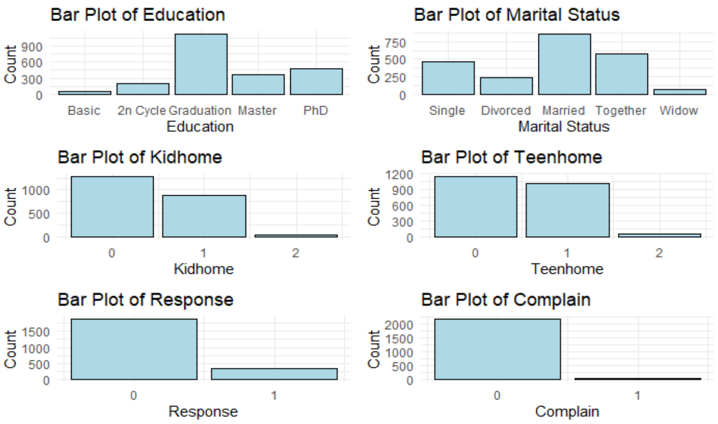

Figure 3 - Bar Plots of Categorical Variables

3. **Outliers:** we identified outliers in 'Year_Birth' and 'Income' variables. These outliers were likely errors or placeholders and were removed from our dataset to maintain data integrity. To identify these outliers, we employed two methods:
   - Cook's Distance: this method identified the outlier in the 'Income' variable but was not capable of identifying the ones for the 'Year_Birth' variable being necessary to use the standard deviation method for such process.

- Standard Deviation Method: using the mean and standard deviation of 'Year_Birth', we identified three outlier values. These entries were removed from our dataset.

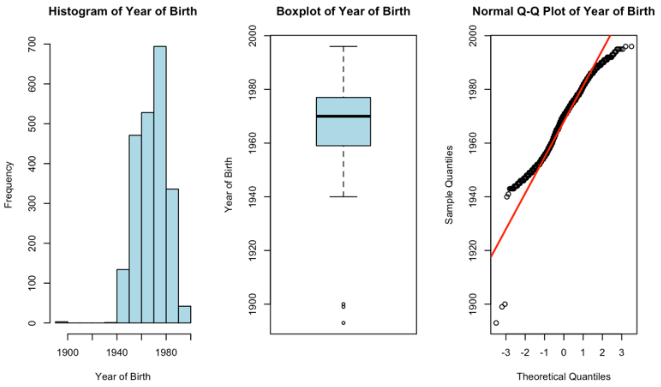

Figure 4 – Year of Birth Outliers

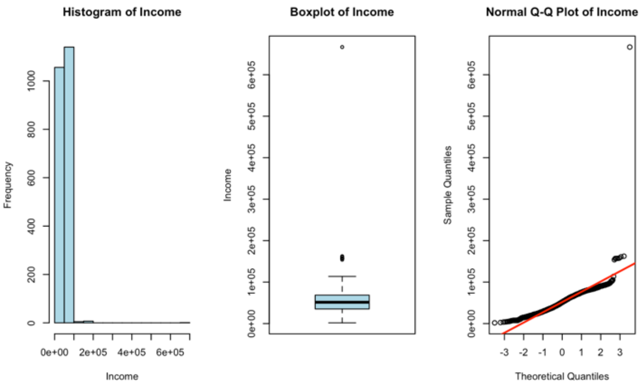

Figure 5 – Income Outliers

4. **Correlation Matrix:** we examined the relationships between different variables. Some significant correlations included a positive correlation between the amount spent by a customer and their income, a negative correlation between having kids at home and both store purchases and catalog purchases, and a positive correlation between income and membership acceptance.

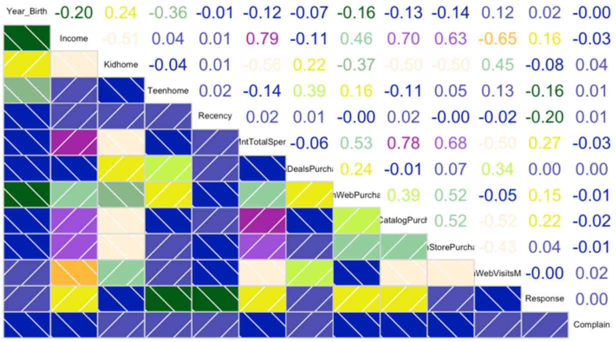

Figure 6 - Correlation Matrix

3.4. Key Variables

Several variables can be identified as key variables for predicting the response:

- **Income:** higher income levels correspond to greater spending. Higher-income customers may prefer shopping in-store or via catalogs.
- **Year of Birth:** to segment our customer base by age.
- **Education:** Most of our customers have graduate-level education. This can influence our marketing approach.
- **Marital Status:** this can affect purchasing behaviors, with differences between single and partnered customers.
- **Number of Kids and Teens at Home:** these impacts purchasing habits and shopping preferences.

- **Total Amount Spent:** to understand factors that influence total spending.

3.5. Initial Hypothesis

Several hypotheses may explain a customer's positive response to a gold membership offer:

1. Higher-income individuals might be more likely to accept the offer, perceiving more value from the membership benefits. This hypothesis is based on the observed positive correlation between income and membership acceptance.
2. Catalog Purchasers: frequent catalog customers may respond positively to the gold membership. These customers already appear engaged and receptive to the superstore's offers, indicating a potential willingness to accept membership.
3. Presence of Kids: customers with kids at home tend to spend more, suggesting a positive correlation between the presence of kids and expenditure. However, they also seem to shop less in-store and through catalogs, possibly finding the convenience and exclusive benefits of membership attractive.

These hypotheses suggest income, catalog purchasing habits, and the presence of kids at home may be significant factors in predicting membership acceptance.

3.6. Feature Engineering

We proceeded with additional data transformations to further optimize our dataset format for subsequent model utilization. Detailed below are the steps taken for this feature engineering phase.

1. **Binary Transformation of Kidhome and Teenhome variables:** only 46 entries show households with two kids at home, and similarly, just 51 entries indicate two teenagers at home. Given this limited representation, we have converted these variables to binary: households with kids (or teens) and without.
2. **Creation of Interaction Terms:** to capture the combined effect of multiple variables, we have created interaction terms between 'Income' and several purchase behavior indicators. The interaction terms we created are Income_NumStore_Interact, Income_NumWeb_Interact, and Income_MntTotal_Spent.

4. Modeling and Hyperparameter Optimization

In the process of building our prediction models, we initially split our dataset into two distinct parts, where 80% of the data served as the training set for building our models, while the remaining 20% was set aside as a testing set for evaluating our models' performance.

1. **Logistic Regression:**

We began with a Logistic Regression model to predict customer responses to a gold membership offer, suitable due to the binary target variable and the model's simplicity. Using all dataset predictors, we built a baseline model classifying 90.48% of instances accurately, and with a ROC curve AUC value of 0.8886, indicating strong differentiation between responses.

To refine our model, we used a stepwise variable selection technique based on Akaike Information Criterion (AIC), progressively removing insignificant predictors from the full model until AIC no longer improved. This created "Model 2" with improved accuracy of around 90.25%.

To ensure model validity, we examined multicollinearity using the Variance Inflation Factor (VIF). Some variables in the first model showed high multicollinearity and were dropped, leading to "Model 3", less susceptible to multicollinearity but with a slightly reduced accuracy of 89.34%. Despite this, we chose Model 2, as it offered a balance between predictive accuracy and model complexity while minimizing multicollinearity.

2. Decision Tree:

The Decision Tree model was developed to predict customer responses to a "Gold Membership" offer in a superstore's discount program. As a tree-based method, often referred to as CART (Classification and Regression Trees), it's applicable across varied contexts, and the final model can be visualized and interpreted easily.

The model began with a minor complexity parameter (cp), building an initial decision tree with all predictor variables. At the end, only eight factors were used: Income, Marital_Status, MntTotalSpent, NumCatalogPurchases, NumDealsPurchases, NumStorePurchases, NumWebVisitsMonth, and Recency. 'MntTotalSpent' resulted in the most important predictor. Testing this tree on the 'Response' variable resulted in an accuracy of 89.12% and a misclassification error rate of 10.88%, demonstrating effective predictive capability with an AUC score of 0.73.

To optimize, we pruned the tree using the cp that minimized cross-validated error, narrowing down to six variables. Though this pruned model showed slightly lesser performance (AUC 0.71, error rate 11.79%), it retained good predictive ability. But, on comparison, the original tree presented superior predictive power.

3. Random Forest:

Our final model was Random Forest, which combines multiple decision trees to produce a more accurate prediction by averaging their outcomes, reducing variance and limiting overfitting. It's a preferred algorithm, especially when dealing with a multitude of variables, due to its strong predictive power.

We first implemented our Random Forest model with all features. The result showed a training accuracy of around 86.22% and an Out of Bag (OOB) error rate of 13.78%. The model's variable importance plot underscored 'MntTotalSpent' and 'Income' as the most significant features. On the test dataset, the initial model displayed a high predictive capacity, correctly classifying about 90.5% of instances with an AUC value of 0.7944.

Optimization of the 'mtry' parameter provided an optimal value of 5, leading to a model with a 14.23% OOB error and 90.0% accuracy and 0.7651 AUC on the test dataset. Compared to this model, the initial model showed a slightly superior AUC score, designating it as our final Random Forest model.

5. Results

5.1. Models Performance and Evaluation

In evaluating the performance of different machine learning models for predicting Gold Membership enrollment, five key metrics were considered: Accuracy, Area Under the Curve (AUC), Sensitivity, Specificity, and Precision. These metrics provide insights into the models' predictive capabilities and their ability to correctly identify positive and negative instances.

Model	Accuracy	AUC	Sensitivity	Specificity	Precision
Logistic Regression	0.9025	0.8897	0.6957	0.9139	0.3077
Decision Tree	0.8912	0.7328	0.5556	0.9210	0.3846
Random Forest	0.9048	0.8821	0.6667	0.9221	0.3846

Figure 7 - Models Performance Comparison

Based on these metrics, the following analysis was conducted:

- **Accuracy:** The Random Forest model achieved the highest accuracy of 90.48%. The Logistic Regression model followed closely with an accuracy of 90.25%.

- **AUC:** We used the AUC (Area Under the ROC Curve) as a crucial metric for predicting Gold Membership enrollment, providing insights into the model's ability to differentiate between positive and negative instances and guiding decision-making for marketing optimization.

  The Logistic Regression model demonstrated the highest AUC score of 88.97%, indicating its good discriminatory power in distinguishing between positive and negative instances. The Random Forest model achieved the second-highest AUC value of 88.21.

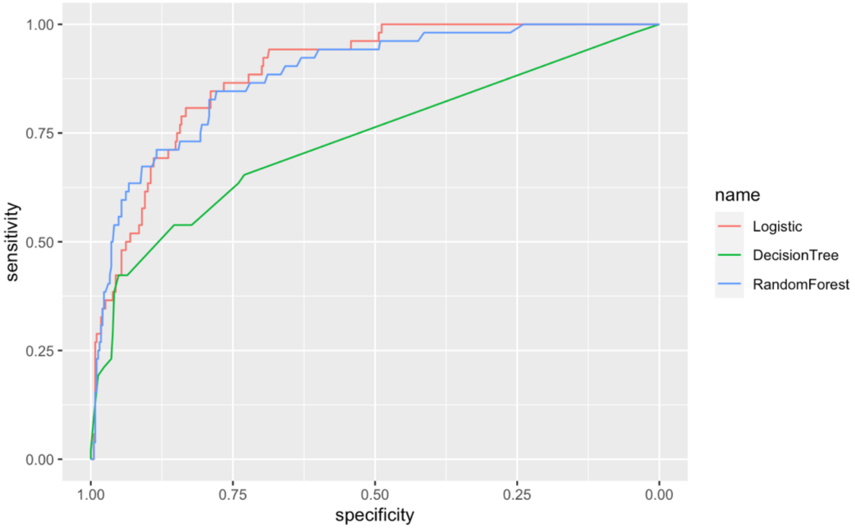

Figure 8 - Models AUC Comparison

- **Sensitivity:** The Logistic Regression model had the highest sensitivity of 69.57%. The Random Forest model followed with a sensitivity of 66.67%, while the Decision Tree model had the lowest sensitivity.
- **Specificity:** The Random Forest model demonstrated the highest specificity value of 92.21%, indicating its ability to correctly identify negative cases in the dataset. The Logistic Regression model had a slightly lower specificity of 91.39%.
- **Precision:** both the Decision Tree and Random Forest models had the highest precision values at 38.46%. The Logistic Regression model had a lower precision value of 30.77%.

5.2. Model Selection

After a detailed examination and comparative analysis of the three machine learning models (Logistic Regression, Decision Tree, and Random Forest), we have concluded that both the Logistic Regression and Random Forest models showed outstanding performance. The Random Forest model slightly outperformed the Logistic Regression model in terms of Accuracy (90.48%), Specificity (92.21%), and Precision (38.46%). However, the Logistic Regression model showed superior performance in terms of AUC (88.97%) and Sensitivity (69.57%), while matching the highest Precision achieved by the Random Forest model.

Although the Random Forest model demonstrated slightly higher accuracy, its relatively lower AUC suggests a weaker discriminatory power between positive and negative instances compared to the Logistic Regression model. On the other hand, the Logistic Regression model's slightly lower accuracy was compensated by its higher AUC, highlighting its superior ability to differentiate between positive and negative instances, which maximizes its predictive efficacy.

In addition, the Logistic Regression model offers better interpretability and simplicity, making it a more favorable choice for our purposes. While the Random Forest model can serve as an excellent performance benchmark, its complexity creates challenges in interpreting and explaining variable interactions.

In conclusion, despite both models performing exceptionally well, the Logistic Regression model, with its strong predictive power and interpretability, has been selected as the primary model for predicting customer responses and Gold Membership enrollment and guide our decision-making and marketing strategies effectively.

5.3. Conclusions and Recommendations

In conclusion, our analysis goal was to develop a predictive model to identify customers with a higher likelihood of accepting the "Gold Membership" offer at the superstore. The Logistic Regression model resulted as the optimal choice, providing valuable insights into customer responses, and influencing factors.

The findings from the model reveal several key insights:

- Higher education levels ('Master's' or 'PhD') significantly increase the likelihood of Gold Membership enrollment, suggesting educated individuals are more inclined to join.
- Marital status: 'Divorced' or 'Widowed' individuals are more likely to enroll, while those 'Together' or 'Divorced' are less likely. Targeting specific customer segments can leverage these findings.
- Newer customers have a higher likelihood of enrolling, indicating the importance of customer retention strategies.
- Increasing 'Recency' (time since last interaction) decreases the likelihood of enrollment, highlighting the need for consistent customer engagement.
- Purchase behavior varies: Deal, web, and catalog purchases positively influence enrollment, while in-store purchases have a negative impact. Online and deal-oriented shopping drive membership enrollment.
- 'NumWebVisitsMonth' (online engagement) positively influences enrollment, emphasizing the importance of a strong online presence and seamless digital experience.
- Teenagers at home negatively impact enrollment, possibly due to different spending priorities in these households.
- Interaction terms 'Income_NumStore_Interact' and 'Income_MntTotal_Spent' show positive coefficients, indicating the interactive effects of income, store interaction, and total spending on enrollment likelihood.

These findings present an opportunity for the superstore to optimize its marketing campaign by targeting the specific segment identified and tailoring incentives to align with their preferences. By strategically focusing on these customers described above on the findings, who exhibit a higher propensity for Gold Membership enrollment, the superstore can enhance customer satisfaction, loyalty, and overall revenue.

To maintain a competitive edge, it is recommended that the superstore regularly updates and refines the model with new data, ensuring its ongoing accuracy in predicting customer responses and supporting future growth in the dynamic retail landscape. By continuously monitoring and adapting to customer behavior, the superstore can stay ahead of the competition and drive sustainable success.

5.4. Further Analysis to Consider

Given the scope of our analysis and the resources available, there are several areas for further exploration and investigation. If provided with additional time and resources, the following areas could be pursued to enhance our understanding and refine the predictive model for Gold Membership enrollment:

1. **Customer Segmentation:** expanding our analysis to conduct a customer segmentation based on demographic and behavioral factors could provide deeper insights into the preferences and characteristics of different customer segments.
2. **Customer Lifetime Value (CLV):** incorporating CLV analysis into our predictive model would enable us to prioritize customers based on their potential long-term value. By considering both the likelihood of Gold Membership enrollment and the expected revenue generated over time, we could tailor retention strategies for high-value customers.
3. **External Data Sources:** incorporating external data sources, such as socioeconomic data, could provide additional context and improve the predictive power of our model.
4. **Experimental Design:** conducting controlled experiments, such as A/B testing, to evaluate the effectiveness of different marketing strategies and incentives could provide more evidence of their impact on Gold Membership enrollment.

5. **Customer Satisfaction and Feedback Analysis:** exploring customer satisfaction metrics and feedback data could provide insights into the drivers of customer loyalty and their perception of the Gold Membership program.

These areas of further analysis have the potential to enhance our understanding of customer behavior, improve the accuracy of our model, and drive more effective marketing strategies for Gold Membership enrollment.

6. References

- James, Gareth, et al. *An Introduction to Statistical Learning with Applications in R*. 2nd ed., Corrected Version, 21 June 2023.
- "Superstore Marketing Campaign Dataset." Kaggle, https://www.kaggle.com/datasets/ahsan81/superstore-marketing-campaign-dataset.
- Deighton, Katie. "Want Better Customer Service? Join the Membership Club." The Wall Street Journal, 2023, https://www.wsj.com/articles/want-better-customer-service-join-the-membership-club-d57725f8.
- McKinsey & Company. "Coping with the Big Switch: How Paid Loyalty Programs Can Help Bring Consumers Back to Your Brand." McKinsey & Company, 2020, https://www.mckinsey.com/capabilities/growth-marketing-and-sales/our-insights/coping-with-the-big-switch-how-paid-loyalty-programs-can-help-bring-consumers-back-to-your-brand.
- Strauss, Valerie. "A Guide to Writing an Academic Paper." The Washington Post, 19 Jan. 2012, https://www.washingtonpost.com/blogs/answer-sheet/post/a-guide-to-writing-an-academic-paper/2012/01/18/gIQAjGCTCQ_blog.html.
- Burns, Emily. "Data Cleaning in R Made Simple." Towards Data Science, 2021, https://towardsdatascience.com/data-cleaning-in-r-made-simple-1b77303b0b17."
- Raoniar, Rahul. "Modelling Binary Logistic Regression using R." OneZero, 2020, https://onezero.blog/modelling-binary-logistic-regression-using-r-research-oriented-modelling-and-interpretation/.F
- Cartaya, Claudia. "Comparison of the Logistic Regression, Decision Tree, and Random Forest Models to Predict Red Wine." Towards Data Science, 2020, https://towardsdatascience.com/comparison-of-the-logistic-regression-decision-tree-and-random-forest-models-to-predict-red-wine-313d012d6953.